УДК 330.43(075.8)

ОЦЕНКА РЕГРЕССИИ С УЧЁТОМ ОГРАНИЧЕНИЙ НА ПАРАМЕТРЫ В УСЛОВИЯХ ГЕТЕРОСКЕДАСТИЧНОСТИ

В.А. Талызин,
Казанский (Приволжский) федеральный университет, г. Казань

Ключевые слова: *оценка, параметры, модель, учёт, ограничения, гетероскедастичность.*

В работах [1;2] рассматривались вопросы оценки коэффициентов моделей с учётом линейных ограничений на них в условиях, когда выполнялись все предпосылки метода наименьших квадратов (МНК). На практике эти условия не всегда выполняются и, в частности, дисперсии ошибок регрессии в разных наблюдениях могут быть различными, т.е. наблюдается гетероскедастичность. Использование обычного МНК в этих условиях приводит к потере эффективности оценок и при малом объёме выборки полученные параметры могут существенно отличаться от оцениваемых.

В этом случае полученные в данных работах методы оценки коэффициентов модели с ограничениями нуждаются в корректировке. Вначале рассмотрим решение задачи на примере линейной парной регрессии.

По исходным данным $(x_i, y_i), i = \overline{1,n}$ требуется оценить линейную парную регрессию

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, i = \overline{1,n}, \quad (1)$$

когда коэффициенты β_j удовлетворяют линейному ограничению

$$\beta_0 + \beta_1 = \alpha, \quad (2)$$

а ошибки регрессии ε_i имеют различные дисперсии $D(\varepsilon_i) \neq const$.

Для оценки коэффициентов модели используем взвешенный метод наименьших квадратов. Поскольку истинные значения $D(\varepsilon_i), i = \overline{1,n}$ неизвестны, то необходимо сделать реалистичные предположения о значениях $D(\varepsilon_i) = \sigma_i^2$. Исследуем два варианта. Пусть вначале дисперсии пропорциональны квадрату фактора x_i:

$$\sigma_i^2 = \sigma^2 x_i^2.$$

Разделим обе части уравнения (1) на известное значение x_i

$$\frac{y_i}{x_i} = \beta_0 \frac{1}{x_i} + \beta_1 + \frac{\varepsilon_i}{x_i}, i = \overline{1,n}$$

и сделаем замену переменных:

$$y_i' = \frac{y_i}{x_i}, \quad x_i' = \frac{1}{x_i}, \quad \varepsilon_i' = \frac{\varepsilon_i}{x_i}, \quad i = \overline{1,n}.$$

Тогда исходное уравнение запишется

$$y_i' = \beta_1 + \beta_0 x_i' + \varepsilon_i', \quad i = \overline{1,n}, \quad (3)$$

в котором свободный член и коэффициент регрессии поменялись местами, а ошибки регрессии ε_i' удовлетворяют условиям гомоскедастичности

$$D(\varepsilon_i') = D(\frac{\varepsilon_i}{x_i}) = \frac{1}{x_i^2} D(\varepsilon_i) = \frac{\sigma^2 x_i^2}{x_i^2} = \sigma^2 = const .$$

Уравнение (3) уже можно оценить обычным МНК

$$\tilde{y}' = b_1 + b_0 x' ,$$

когда параметры b_0, b_1 должны удовлетворять условию

$$b_0 + b_1 = \alpha . \qquad (4)$$

Из уравнения (4) выразим

$$b_1 = \alpha - b_0 \qquad (5)$$

и вставим в выражение для остаточной суммы квадратов

$$Q_e = \sum_{i=1}^{n} \left((y_i' - b_1 - b_0 x_i')(y_i' - b_1 - b_0 x_i') \right) = \sum_{i=1}^{n} \left((y_i' - b_0 x_i' - (\alpha - b_0))(y_i' - b_0 x_i' - (\alpha - b_0)) \right).$$

Теперь требуется найти безусловный минимум функции Q_e. Приравнивая производную

$$\frac{dQ_e}{db_0} = 2\sum_{i=1}^{n} \left(y_i'(1 - x_i') - \alpha(1 - x_i') + b_0 (1 - x_i')^2 \right)$$

нулю, окончательно получаем формулу для вычисления параметра b_0:

$$b_0 = \frac{\sum_{i=1}^{n} (1 - x_i')(\alpha - y_i')}{\sum_{i=1}^{n} (1 - x_i')^2} . \qquad (6)$$

Далее определяется значение параметра b_1 из формулы (5).

Исследуем случай, когда дисперсии ошибок регрессии пропорциональны фактору x_i:

$$\sigma_i^2 = \sigma^2 x_i .$$

Повторяя аналогичные выкладки с весом $\frac{1}{\sqrt{x_i}}$, нетрудно получить преобразованное исходное уравнение регрессии в виде

$$y_i' = \beta_0 x_{1i}' + \beta_1 x_{2i}' + \varepsilon_i' ,$$

где $y_i' = \frac{y_i}{\sqrt{x_i}}$, $x_{1i}' = \frac{1}{\sqrt{x_i}}$, $x_{2i}' = \sqrt{x_i}$, $\varepsilon_i' = \frac{\varepsilon_i}{\sqrt{x_i}}$, $i = \overline{1,n}$.

Отметим, что преобразованное уравнение представляет теперь двухфакторную множественную регрессию без свободного члена.

Используя соотношение (5), вновь получаем выражение остаточной суммы квадратов как функции только переменной b_0. После приравнивания производной нулю, получаем формулу для вычисления параметра b_0:

$$b_0 = \frac{\sum_{i=1}^{n} (x_{1i}' - x_{2i}')(y_i' - \alpha \cdot x_{2i}')}{\sum_{i=1}^{n} (x_{1i}' - x_{2i}')^2} . \qquad (7)$$

Проиллюстрируем предложенный метод на числовом примере.

Имеются статистические данные с наличием гетероскедастичности, представленные в таблице:

y_i	x_i	y_i	x_i
8,1	6	88,3	95
4,4	3	90,5	79
20,8	18	122	112
12,9	8	132,4	106
28,8	23	114,2	125
15,5	20	99,1	115
48,7	49	156,1	149
37,5	39	150,6	132
104,6	74	362,9	282
68,6	60	209,5	157

Построить линейную модель регрессии $\tilde{y} = b_0 + b_1 x$ по этим данным, когда параметры удовлетворяют условию

$$b_0 + b_1 = 3. \qquad (8)$$

После реализации описанного метода получаем следующие результаты. Если предположить, что $\sigma_i^2 = \sigma^2 x_i^2$, то выборочное уравнение регрессии имеет вид

$$\tilde{y} = 1{,}931 + 1{,}069x.$$

При этом остаточная сумма квадратов составит величину $Q_e = 7536{,}4$.

Без учёта ограничений (8) уравнение регрессии запишется $\tilde{y} = 1{,}382 + 1{,}092x$ с $Q_e = 6837{,}1$. Как и следовало ожидать, дополнительные ограничения на параметры приводят к увеличению остаточной суммы квадратов.

В другом случае, когда $\sigma_i^2 = \sigma^2 x_i$ уравнение регрессии примет вид

$$\tilde{y} = 1{,}887 + 1{,}113x$$

с $Q_e = 6277{,}6$. Без учёта ограничений (8) оценкой будет $\tilde{y} = 0{,}604 + 1{,}069x$ с $Q_e = 6002{,}3$.

Обобщим полученный результат на случай множественной регрессии. Пусть по имеющейся многомерной выборке $(y_i, x_{1i}, ..., x_{pi})$, $i = \overline{1,n}$ требуется получить уравнение регрессии

$$\tilde{y} = b_0 + b_1 x_1 + ... + b_p x_p, \qquad (9)$$

когда выполняются ограничения:

$$\begin{gathered} \alpha_{10} b_0 + \alpha_{11} b_1 + \cdots + \alpha_{1p} b_p = c_1, \\ \alpha_{20} b_0 + \alpha_{21} b_1 + \cdots + \alpha_{2p} b_p = c_2, \\ \cdots \\ \alpha_{m0} b_0 + \alpha_{m1} b_1 + \cdots + \alpha_{mp} b_p = c_m. \end{gathered} \qquad (10)$$

Допустим, что уравнения (10) являются линейно-независимыми, $m < p+1$ и базисный минор матицы коэффициентов системы находится слева, т.е. в первых m столбцах.

Предположим, что по фактору x_s, $1 \le s \le p$, с использованием теста Голдфельда-Квандта обнаружено явление гетероскедастичности. Пусть $\sigma_i^2 = \sigma^2 x_{si}^2$. Тогда после введения переменных

$$y_i' = \frac{y_i}{x_{si}}, \quad x_{0i}' = \frac{1}{x_{si}}, \quad x_{1i}' = \frac{x_{1i}}{x_{si}}, ..., x_{pi}' = \frac{x_{pi}}{x_{si}} \tag{11}$$

получим преобразованное уравнение регрессии

$$y_i' = b_0 x_{0i}' + b_1 x_{1i}' + ... + b_{s-1} x_{s-1i}' + b_s + b_{s+1} x_{s=1i}' + ... + b_p x_{pi}', \tag{12}$$

в котором свободным членом является уже параметр b_s.

Сформируем следующие матрицы и векторы:

$$X_1 = \begin{pmatrix} x_{01}' & x_{11}' & \cdots & x_{s-11}' & 1 & x_{s+11}' & \cdots & x_{p1}' \\ x_{02}' & x_{12}' & \cdots & x_{s-12}' & 1 & x_{s+12}' & \cdots & x_{p2}' \\ . & . & . & . & . & . & . & . \\ x_{0n}' & x_{1n}' & \cdots & x_{s-1n}' & 1 & x_{s+1n}' & \cdots & x_{pn}' \end{pmatrix}, \quad A = \begin{pmatrix} \alpha_{10} & \alpha_{11} & \cdots & \alpha_{1p} \\ . & . & \cdots & . \\ \alpha_{m0} & \alpha_{m1} & \cdots & \alpha_{mrp} \end{pmatrix}, \tag{13}$$

$$Y_1 = \begin{pmatrix} y_1' \\ y_2' \\ \vdots \\ y_n' \end{pmatrix}, \quad b = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}, \quad c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_r \end{pmatrix}. \tag{14}$$

Т

огда уравнение (11) в матричной форме запишется

$$Y_1 = X_1 b,$$

а остаточная сумма квадратов и ограничения (10) примут соответственно вид

$$Q_e = (Y_1 - X_1 b)'(Y_1 - X_1 b),$$

$$Ab = c. \tag{15}$$

Предположим, что $s < r - 1$. Тогда введём в рассмотрение матрицы:

$$X_{11} = \begin{pmatrix} x_{01}' & \cdots & x_{s-11}' & 1 & x_{s+11}' & \cdots & x_{m-11}' \\ x_{02}' & \cdots & x_{s-12}' & 1 & x_{s+12}' & \cdots & x_{m-12}' \\ . & \cdots & . & . & . & \cdots & . \\ x_{0n}' & \cdots & x_{s-1n}' & 1 & x_{s+1n}' & \cdots & x_{m-1n}' \end{pmatrix}, \quad X_{12} = \begin{pmatrix} x_{m1}' & x_{m+11}' & \cdots & x_{p1}' \\ x_{m2}' & x_{m+12}' & \cdots & x_{p2}' \\ . & . & \cdots & . \\ x_{mn}' & x_{m+1n}' & \cdots & x_{pn}' \end{pmatrix},$$

$$A_1 = \begin{pmatrix} \alpha_{10} & \alpha_{11} & \cdots & \alpha_{1m-1} \\ \alpha_{20} & \alpha_{21} & \cdots & \alpha_{2m-1} \\ . & . & \cdots & . \\ \alpha_{m0} & \alpha_{m1} & \cdots & \alpha_{mm-1} \end{pmatrix}, \quad A_2 = \begin{pmatrix} \alpha_{1m} & \alpha_{1m+1} & \cdots & \alpha_{1p} \\ \alpha_{2m} & \alpha_{2m+1} & \cdots & \alpha_{2P} \\ . & . & \cdots & . \\ \alpha_{mm} & \alpha_{mm+1} & \cdots & \alpha_{rp} \end{pmatrix}, \quad b^1 = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_{r-1} \end{pmatrix}, \quad b^2 = \begin{pmatrix} b_r \\ b_{r+1} \\ \vdots \\ b_p \end{pmatrix}.$$

Тогда из векторного уравнения (15) находим

$$b^1 = A_1^{-1}(c - A_2 b^2). \tag{16}$$

После представления $X_1 b = X_{11} b^1 + X_{12} b^2$ получим выражение остаточной суммы квадратов через вектор b^2:

$$Q_e = \left(Y_1 - X_{11} A_1^{-1} c - D b^2\right)' \left((Y_1 - X_{11} A_1^{-1} c - D b^2\right),$$

где

$$D = X_{12} - X_{11} A_1^{-1} A_2.$$

Взяв производную и приравняв её нулю, получаем формулу для вычисления вектора b^2:

$$b^2 = (D'D)^{-1} D'\left(X_{11}A_1^{-1}c - Y_1\right). \qquad (17)$$

Далее по формуле (16) определяется вектор b^1.

Если гетероскедастичность обнаружена по двум факторам x_s и x_l, то предполагая, например, что зависимость дисперсии ошибок регрессии пропорциональными квадрату линейной комбинации этих факторов $\sigma_i^2 = \sigma^2 (x_s + \gamma \cdot x_l)^2$, можно преобразовать уравнение (9) с весами $\dfrac{1}{x_s + \gamma x_l}$ и применить для оценки параметров взвешенный метод наименьших квадратов. Варьируя коэффициентом γ можно выбрать модель с наименьшей остаточной суммой квадратов отклонений.

Рассмотрим случай, когда ограничения на параметры заданы в виде неравенств. Пусть по имеющейся выборке требуется получить уравнение регрессии (9), когда параметры удовлетворяют линейным неравенствам:

$$\begin{gathered} \alpha_{10}b_0 + \alpha_{11}b_1 + \cdots + \alpha_{1p}b_p \ge c_1, \\ \alpha_{20}b_0 + \alpha_{21}b_1 + \cdots + \alpha_{2p}b_p \ge c_2, \\ \cdots \\ \alpha_{m0}b_0 + \alpha_{m1}b_1 + \cdots + \alpha_{mp}b_p \ge c_{mr}. \end{gathered} \qquad (18)$$

Пусть также по фактору x_s, $1 \le s \le p$, с использованием теста Голдфельда-Квандта обнаружено явление гетероскедастичности. Введя в рассмотрение новые переменные (11), получим уравнение регрессии в виде выражения (12). Сформируем матрицы и векторы по формулам (13), (14).

Условия Куна-Таккера запишутся в виде

$$u = 2Qb - A'v + c', \quad s = Ab - q, \quad u'b + s'v = 0 \qquad (19)$$

$$b \ge 0, \quad u \ge 0, \quad v \ge 0, \quad s \ge 0, \qquad (20)$$

где матрица Q имеет вид $Q = X_1'X_1$.

Эффективным и простым методом решения задачи (19)-(20) является метод решения задач о дополнительности [3]. Правило решения этой задачи сформулированы в работе [1].

Список литературы

1. *Талызин В.А.* Оценка параметров эконометрической модели с учётом их ограничений // Вестник КГФЭИ. 2011. № 4. С. 23–26.

2. *Талызин В.А., Кирпичников А.П.* Оценивание параметров эконометрической модели с учётом линейных ограничений // Вестник технологического ун-та. 2015. Т 18. № 13. С. 185–189.

3. *Рейклетис Г., Рейвиндран А., Рэгсдал К.* Оптимизация в технике. В 2-х книгах. М.: Мир, 1986. 774 с.